DOCUMENT RESUME

ED 361 865 EA 025 242

AUTHOR Franklin, Bobby J.; And Others
TITLE Measures of School Effectiveness: Consistency/Inconsistency When Models Are Varied across Subject Area Tests.
PUB DATE Apr 93
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Effective Schools Research; *Evaluation Criteria; High Schools; Models; *Multiple Regression Analysis; Public Schools; *Regression (Statistics); *School Effectiveness; Social Indicators; Validity

ABSTRACT

This paper presents findings of a study that sought to determine whether a composite score is a more appropriate dependent variable in the measurement of school effectiveness than any one component score. The viability of using a composite score, such as the School Effectiveness Index (SEI), was compared with the viability of using any one of five different subject areas. Specifically, the study examined the level of consistency that exists when school-effectiveness-classification models are held constant across test scores. Multiregression analysis of data from 315 Louisiana high schools was used to predict student achievement on a statewide-administered criterion-referenced test (CRT) from indices of socioeconomic status (SES), racial makeup, school report card variables, and demographic data. Findings indicate that a composite index incorporates the varied information that each component issue provides and presents a better overall picture of school effectiveness than would any one test component. In conclusion, a school's success in one subject does not necessarily ensure success in other areas. The composite score is preferable for school-effectiveness research, because it provides more information and is more reliable than a subtest. Two tables and one figure are included. (LMI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 361 865

MEASURES OF SCHOOL EFFECTIVENESS: CONSISTENCY/INCONSISTENCY WHEN MODELS ARE VARIED ACROSS SUBJECT AREA TESTS

Bobby J. Franklin
Linda J. Crone
Arthur M. Halbrook
Louisiana Department of Education
Baton Rouge, Louisiana
and
Michael H. Lang
Louisiana Systemic Initiatives Program
Baton Rouge, Louisiana

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. Franklin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EA 025 242

A paper presented at the 1993 annual meeting of the American Educational Research Association, April, 1993, Atlanta, GA

MEASURES OF SCHOOL EFFECTIVENESS: CONSISTENCY/INCONSISTENCY WHEN MODELS ARE VARIED ACROSS SUBJECT AREA TESTS

In 1991, the Louisiana Department of Education implemented its first school incentive program to acknowledge and reward those public schools which demonstrated progress toward effectively educating their students. The method of determining which schools received awards was based partly on standardized test scores as applied to school category groups. The highest scoring schools in each category on various tests and other indicators received both monetary and nonmonetary awards. In addition, the Department had already begun a school performance comparison program, also based on school category groups. Again, the comparisons were based partly on standardized test scores.

As some of the school indicators of success, the incentive awards and school comparison programs employed the results from the state's existing testing program. Such has been the general practice used in isolating effective schools, both in research and in practice (Good & Brophy, 1986; Weitman, Garber, Oescher, & Brooks, 1990).

The effective schools research began in the 1970s with the hypothesis by some educators that schools could educate their populations regardless of the background of children that they serve (Edmonds & Freiderikson, 1979). The research movement built momentum in the 1980s as definitions of effective schools and various methods of isolating such schools began to emerge. With these definitions came improved methods of comparing schoolwide achievement, including school categorization and regression analysis.

Though much has been written about effective schools, no one method of classifying schools on the criterion of effectiveness has yet to find universal acceptance (Good & Brophy, 1986; Levine & Lezotte, 1990; Purkey & Smith, 1983; Rowan, Bossert, & Dwyer, 1983). Even the widely applied regression model has its problems (Mandeville & Anderson, 1987; Purkey & Smith, 1983; Rowan et al., 1983).

Research on techniques of isolating the effectiveness of individual schools has evolved over a two-decade period since the Dyer, Linn, and Patton (1969) study had attempted to control for student background variables with the regression model. Within that time frame, researchers conducted numerous studies on effective schools, employing various techniques of which the regression model was most frequently used (Lang, 1991). Mandeville and Heidari (1988) also concluded that the regression model was the most frequently employed model in effective school research.

Although the most commonly used model for school effectiveness research appears to be the regression model, an additional controversial issue has emerged as to the choice of indicators used and the stability and consistency of those indicators. Concerned with the stability problem in school effectiveness classifications, Rowan et al. (1983) reviewed existing effective schools research, finding several problematic trends. The authors concluded that presently employed school evaluation procedures were problematic for several reasons: (1) evaluation focused solely on basic skills outcomes, (2) the available procedures for assessing school quality were problematic, and (3) the current procedures presented an incomplete view of school outcomes. In addition, they viewed the selection of test instruments as a potential source of instability:

> In our view, the instability of current measures may result from the fact that they are based on standardized achievement tests that do not accurately reflect the curriculum of the school. Alternative assessment tests that are more closely aligned to the curriculum exist, and practitioners may wish to use these to evaluate the instructional effectiveness of schools. (p. 30)

In addition, their study noted the problems that a narrow focus posed for evaluation procedures: (1) quantitative indicators did not correspond to qualitative conclusions, and (2) limits were placed on the breadth of school improvement programs. That is, what made schools effective extended further than whatever behaviors were measured with basic skills tests. Their findings raised questions as to the validity of narrow measures of student performance when applied to establishing school effectiveness. Finally, the researchers determined that effective school studies seldom measured instructional performance from available data on grades and subject areas. Such limitations confounded the stability issue in those studies.

Underlying this issue of stability is another issue—consistency. Consistency is defined in this study as the ability of a regression model to accurately isolate effective and ineffective schools at one point in time. Therefore, random and systematic errors in school effectiveness classifications are major concerns of consistency. Conceptually, consistency is a necessary, but not a sufficient condition of stability. Without consistency, there can be no stability; however, consistency is no guarantee that stability will exist.

Consistency can be further defined as that quality which allows the model to produce similar results under different situations at the same general point in time (Lang, 1991).

By varying the situation such as using parallel test forms or systematically splitting populations, the quality is threatened in terms of degree. If the resulting classifications are inconsistent, then the stability of such results from one point in time to the next is in jeopardy.

Complications associated with model variations on school effectiveness classifications were noted by Levine and Lezotte in a 1990 monograph on effective schools:

> Researchers who have carefully examined the data in school effectiveness studies generally have concluded that many schools identified as particularly successful according to a particular measure such as reading scores or sub scores at a particular grade do not stand out as unusually successful with respect to other grade levels, other subject areas, and alternate performance measures (norm-referenced or criterion-referenced) in the same subject or related area. (p. 4)

Levine and Lezotte (1990) suggested caution in drawing conclusions from effective school research findings, noting that varying the achievement criteria often influenced resulting classifications. However, such findings are presently providing the foundation for state and district school evaluation and incentive award programs (Weitman et al., 1990).

Many other researchers have expressed concern with the use of just one subject area or grade level as the school effectiveness index. Purkey and Smith (1983) felt that using only one subject area and grade level as an index gives a very limited view of a school's effectiveness. Witte and Walsh (1990) found that different variables were needed to predict the different subject areas of reading and mathematics. Mandeville has conducted

numerous studies on comparing the consistency and stability when using different subject areas and grade levels as the dependent variables (Mandeville, 1987; Mandeville & Anderson, 1987; Mandeville, 1988). He reported finding that the strongest cross-year consistency of scores was with a combined reading-mathematics score (Mandeville, 1987), which was computed by averaging the two subject areas.

Consequently, it appears that one solution to this controversy may be to use some type of composite of different subject areas and grade levels. This study investigates the viability of using a composite score as the School Effectiveness Index (SEI) compared to using any one of five different subject areas. The level of consistency that exists when school effectiveness classification models are varied along test scores will be examined. The school classification models are held constant across school input variables, but are varied across school output variables (e.g., test scores). Five different subject areas (English language arts, mathematics, written composition, science, and social studies), will be used as the output variables, as well as a composite of all five subjects.

The research questions examined within this study were (1) whether a model employing the results of a singular test produced consistent results with a model using the composite test scores, and (2) whether various models employing singular test scores produced consistent results with one another.

Methodology

Sampling

As the intent was to compare as many subject areas as possible to a composite, it was deemed most appropriate to use schools containing both grades 10

and 11 for this study. In Louisiana, the Graduation Exit Examination (GEE), which is administered to students in grades 10 and 11, consists of English language arts, mathematics, and written composition at grade 10, and science and social studies at grade 11. In order to make the study as generalizable as possible, all regular education schools that contained grades 10 and 11 throughout the state were included in the study. This provided a good mixture of metropolitan, suburban, and rural schools, incorporating many sizes and many socioeconomic levels. Eliminated from the study were alternative schools, P.M. schools, and schools with missing data on any variable. The final sample included 315 schools, with the school being the unit of analysis.

Classification Model

The schools were classified into three categories (effective, average, and ineffective), utilizing a multiple regression analysis, predicting student achievement on a statewide administered criterion-referenced test (CRT) from indices of socioeconomic status (SES), racial makeup, school report card variables, and demographic data. The indicator of SES was the percent of students participating in the free lunch program and racial makeup consisted of the percent of Afro-American students in the school. School report card variables included the percent of teachers with a Ph.D. in each school, as well as the output variables of percent attendance and percent suspension. Community type and the interaction between community type and percent of students on free lunch were also used as predictor variables.

Measures Used to Select Schools

The school report card indicators used in the regression were collected for the Louisiana Progress Profiles Program, which produces a yearly report of a school's "health" for each regular education school throughout the state of Louisiana. The data reported in this school report card were collected from many bureaus in the Louisiana Department of Education, but underwent extensive correction and verification procedures before being used by the Progress Profiles Program.

The SES measure and racial makeup used in the regression are collected as part of the statewide testing program. The students report their race and participation in the free lunch program while taking the test.

Dependent Variables

The CRTs are administered as part of the Louisiana Educational Assessment Program, which tests grades three, five, seven, ten and eleven in all public schools throughout the state. This test was produced by the Louisiana Department of Education (LDE) and was developed to measure the attainment of the curricular guidelines specified by the state. The CRTs are not minimum skills test, but are designed to measure grade level skills. The state's curriculum guides are constructed with specific standards for each grade level and subject area, and the CRT items are then designed and validated in order to reflect those standards. The GEE, which is administered in grades 10 and 11, is part of this program (Louisiana Department of Education, 1989).

In order to make comparisons of classification using different subject areas as the dependent variable to a classification using the composite score as the dependent variable, it was deemed necessary to transform all student level scores to scores with a common mean and standard deviation (Hinkle, Wiersma, & Jurs, 1988, chap. 4). Hence, the student level scores were transformed to t-scores (mean of 50, standard deviation of 10) with the use of the state population means and standard deviations for each separate grade level and subject area. The school level composite scores were then computed by averaging the student level t-scores for all five subject areas.

Procedure

With the use of the regression model that best fit the composite score (percent Afro-American, percent attendance, percent of teachers with a Ph.D., percent of students receiving free lunch, percent of students suspended, community type, and interaction of community type and free lunch), separate regression procedures were run for each subject area, as well as for the composite. When using the composite score as the dependent variable, this model yielded an R^2 of .61 ($p<.0001$). The residual for each school was used to label the schools as effective, average, or ineffective, with cutoff points of +/-.674 standard deviations. Recent research (Lang, Teddlie, & Oescher, 1992) indicates that cutoff points at this level minimize the effect of expected chance consistency in comparing school effectiveness models. After classification of schools into the three school effectiveness categories (effective, average, and ineffective) for each subject area and the composite, consistency of

classification was compared, making comparisons of each individual subject to the composite and each subject area to the other.

Two types of analysis were conducted: (a) Relationships of the subject areas to the composite were examined utilizing correlations of the composite and subject area t-scores, and correlations of the composite and subject area residuals; and (b) consistency of classification of school effectiveness categories was compared using an unweighted agreement ratio and an unweighted kappa coefficient (an agreement ratio that controls for chance agreement expected from the distribution of the data). For the agreement ratio and the kappa coefficient, the data was placed in a 3-by-3 contingency table (See Figure 1.), and the percent of those schools that were consistently classified was computed.

		SCHOOL CLASSIFICATION BY COMPOSITE		
		Effective	Average	Ineffective
SCHOOL CLASSIFICATION BY MATHEMATICS	Effective	EE	EA	EI
	Average	AE	AA	AI
	Ineffective	IE	IA	II

Figure 1: A example of a 3 x 3 contingency table used in the calculation of agreement ratios and kappa coefficients.

The agreement ratio was then computed by summing the percent of schools that fall in the consistent classification cells of effective/effective, average/average, or ineffective/ineffective (i.e., the diagonal). Employing the table's row and column totals

(marginals) to determine expected agreement, the kappa coefficient adjusted the agreement ratio for expected chance agreement.

Findings

The primary issue of this study was whether a composite score would be a more appropriate dependent variable in the measurement of school effectiveness than would any one component score.

The agreement indices found in Table 1 demonstrate that there is limited or little consistency in school effectiveness classifications when individual component scores are employed (.600-.498 for the agreement ratio; .366-.162 for the kappa coefficient). However, stronger consistency is found when the individual components are compared to a composite score (.679-632 for the agreement ratio; .490-.396 for the kappa coefficient).

The consistency with which the composite scores were able to classify schools in accordance with individual component scores was approximately 6.5 of every 10 schools using the agreement ratio, and 4 to 5 of every 10 schools using the kappa coefficient. Those consistency ratios were less for the component scores when compared with one another: approximately 5 to 6 of every 10 schools using the agreement ratio, and 1.5 to 3.6 of every 10 schools using the kappa coefficient.

Obviously, the composite score is influenced by each of the component scores. Since it is computed from a combination of each component, the composite can generally be expected to have a stronger relationship to each component than do the components to each other.

Table 1

Correlations, agreement ratios, and kappa coefficients for the composite and subject area component SEI scores.

	Composite	Language Arts	Math	Science	Social Studies
Composite					
Language Arts	r= 0.797 $\%A_U$= 0.679 K_U= 0.487				
Math	r= 0.725 $\%A_U$= 0.679 K_U= 0.490	r= 0.584 $\%A_U$= 0.600 K_U= 0.366			
Science	r= 0.680 $\%A_U$= 0.638 K_U= 0.404	r= 0.368 $\%A_U$= 0.508 K_U= 0.194	r= 0.318 $\%A_U$= 0.517 K_U= 0.216		
Social Studies	r= 0.732 $\%A_U$= 0.632 K_U= 0.396	r= 0.414 $\%A_U$= 0.524 K_U= 0.222	r= 0.308 $\%A_U$= 0.498 K_U= 0.188	r= 0.638 $\%A_U$= 0.578 K_U= 0.282	
Written Composition	r= 0.605 $\%A_U$= 0.651 K_U= 0.431	r= 0.514 $\%A_U$= 0.587 K_U= 0.331	r= 0.250 $\%A_U$= 0.514 K_U= 0.218	r= 0.153 $\%A_U$= 0.502 K_U= 0.162	r= 0.273 $\%A_U$= 0.505 K_U= 0.172

A review of the Pearson correlation coefficients for the average test scores for each school (i.e., the dependent variables for the regression models) provides additional evidence as to why the composites have a stronger relationship. Composite scores in general can be expected to relate higher to the component scores than the components to each other. The data in Table 2 demonstrate that very phenomenon. One reason for the phenomenon is that the information from each component constitutes a part of the composite, thus enhancing its correlation with the composite.

Another is that each component should provide unique information on a school which can be expected to limit the degree of correlation among components.

Additionally, the data in Table 2 demonstrate that the language arts and mathematics scores have a moderate relationship with each other as do science and social studies. But when those test results are crossed (i.e., language arts and science or social studies, mathematics and science or social studies), the relationships are considerably weaker. A possible non-content explanation for this phenomenon is that the language arts and mathematics tests were administered to the 10th grade population for 1992, whereas the science and social studies tests were administered to the 11th grade population for that same year. Therefore, the additional information being provided beyond content issues may be due to different populations.

Furthermore, scores on the written composition measure demonstrate little consistency with not only the Grade 11 tests but also the Grade 10 mathematics test. The written composition instrument is the only performance test administered in the GEE. Though its results demonstrate a strong relationship with the language arts test, other GEE tests with less related content orientation demonstrate a more moderate relationship. That phenomenon suggests that the results of the written composition performance test are providing substantial unique information about student performance that many of the other instruments do not.

In summary, the agreement indices suggest to the presenters that a composite index will incorporate the varied information which each component index provides and give a better overall picture of school effectiveness than would any one test component.

Table 2

Consistency Measures in Rank Order Format

Rank	Pearson School Scores	Pearson SEIs	Unweighted Agree Ratio	Unweighted Kappa Coef
1	COMP/LA .918	COMP/LA .767	COMP/LA .679	COMP/MT .490
2	COMP/MT .853	COMP/SS .732	COMP/MT .679	COMP/LA .487
3	COMP/SC .826	COMP/MT .725	COMP/WC .651	COMP/WC .431
4	COMP/SS .826	COMP/SC .680	COMP/SC .638	COMP/SC .404
5	COMP/WC .786	SC/SS .638	COMP/SS .632	COMP/SS .396
6	LA/MT .784	COMP/WC .605	LA/MT .600	LA/MT .366
7	SC/SS .771	LA/MT .584	LA/WC .587	LA/WC .331
8	LA/WC .739	LA/WC .514	SC/SS .578	SC/SS .282
9	LA/SC .676	LA/SS .414	LA/SS .524	LA/SS .222
10	LA/SS .662	LA/SC .368	MT/SC .517	MT/WC .218
11	MT/SC .599	MT/SC .318	MT/WC .514	MT/SC .216
12	MT/SS .569	MT/SS .308	LA/SC .508	LA/SC .194
13	MT/WC .561	WC/SS .273	WC/SS .505	MT/SS .188
14	WC/SS .540	MT/WC .250	WC/SC .502	WC/SS .172
15	WC/SC .498	WC/SC .153	MT/SS .498	WC/SC .162

Discussion

The results of this study substantiate the beliefs of other researchers that a school's success in one subject area does not necessarily assure that they are successful in other areas (Levine & Lezotte, 1990). When examining the agreement ratios of the different subject areas, the strongest agreement (not controlling for chance) was between English Language Arts and Mathematics (.60) which means that 4 out of 10 schools would not be classified the same (effective, average, or ineffective) if the two different subject areas were used as the school effectiveness index. When comparing some subject areas, this agreement ratio drops to the point that approximately 50% of the schools would be inconsistently classified if different subject areas were used as the dependent variable. When taking chance classification into consideration, the percent of agreement between some components is extremely low (.162 for classification using written composition compared to classification using science). In other words, with chance agreement controlled for, less than 2 out of 10 schools would be classified consistently if written composition were used instead of science. The correlations of the residuals provide a similar pattern of relationships of the different subject areas to each other and stronger relationships of the subject areas to the composite.

Since English Language Arts yields the highest relationship of any subject area to the composite and also relates higher than any other subject area to both mathematics and written composition, this subject appears to be the best choice if only one subject were available for use in a study. This still gives a very limited view

of the effectiveness of a school. Just as there is criticism that academic achievement provides a limited view of school effectiveness (Weitman, et al., 1990), one subject area and grade level also provides a very narrow perspective of academic achievement.

This study indicates that the composite is a preferable score for use in school effectiveness research. This confirms the findings of Mandeville (1987) who reported that a combined reading-mathematics score produced results which were more consistent within grades and across years than the individual subject area scores. Considering that the composite is a product of all the subject areas, it obviously provides more information than any single subject score. It appears that written composition especially provides new information to the composite. Its low correlations and agreement ratios when compared to the other subject areas indicates very little shared variance with any subject area except English Language Arts. Written Composition, being a performance based test, may be measuring other dimensions of academic performance. The other components of the test require multiple choice responses. The addition of written composition responds to the concern of Rowen, et al. (1983) that additional measures besides basic skills tests should be included.

In addition to providing more information about a school, the composite also is reported to be a more reliable measure than a subtest (Crocker & Algina, 1986). Two factors cited as providing an increase in decision consistency of a test score are test length and test score generalizability. Obviously, the composite score consists of a

greater number of items and a greater sampling of knowledge than any of the subject area components.

Further research is recommend, both in different geographical areas and using different tests. Also, it would be advisable to examine the consistency across years for these different subject areas compared to the composite score. We do recommend that component scores be transformed to z- or t-scores prior to averaging for a composite. If the composite were found to be more consistent from one year to the next (as Mandeville reported that the combined language arts-mathematics score was), this would provide an even stronger argument for its use.

References

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Dyer, H.S., Linn, R.L., & Patton, M.J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 6(4), 591-605.

Edmonds, R.R. & Frederiksen, J.R. (1979). Search for effective schools: The identification and analysis of city schools that are instructionally effective for poor children. Washington, DC: US Office of Health, Education, and Welfare.

Good, T.L. & Brophy, J. (1986). School effects. In M. Wittrock, Ed., Third handbook of research in teaching. New York: Macmillan.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1988). Applied statistics for the behavioral science. Boston: Houghton Mifflin Company.

Lang, M.H. (1991). Effective school status: A methodological study of classification consistency. Unpublished doctoral dissertation, Louisiana State University, Baton Rouge.

Lang, M.H., Teddlie, C., & Oescher, J. (April, 1992). The effect that varying the test mode had on school effectiveness ratings. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Levine, D.U. & Lezotte, L.W. (1990). Unusually effective schools. Madison, WI: The National Center for Effective Schools Research and Developement.

Louisiana Department of Education (1991). Louisiana progress profile: Summary school reports, 1989-1990. Baton Rouge: Author.

Mandeville, G.K. (April, 1987). On the identification of effective and ineffective schools. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Mandeville, G.K. (1988). School effectiveness indices revisited: Cross-year stability. Journal of Educational Measurement, 25(4), 349-356.

Mandeville, G.K. & Anderson, L.W. (1987). The stability of school effectiveness indices across grade levels and subject areas. Journal of Educational Measurement, 24(3), 203-216.

Mandeville, G.K. & Heidari, K. (April, 1988). Measuring school effectiveness using hierarchical linear models. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Purkey, S.C. & Smith, M.S. (1983). Effective schools: A review. The Elementary School Journal, 83(4), 427-452.

Rowan, B., Bossert, S.T., & Dwyer, D.C. (1983). Research on effective schools: A cautionary note. Educational Researcher, 12(4), 24-31.

Weitman, C.J., Garber, D.H., Oescher, J., & Brooks, C. (1990). Louisiana incentive program: Policies and recommendations 1990. Report for the Louisiana Department of Education. Ruston: Louisiana Tech University.

Witte, J.F., & Walsh, D.J. (1990). A systematic test of the effective schools model. Educational Evaluation and Policy Analysis, 12(2), 188-212.